Risk Analysis, Vol. 12, No. 1, 1992

Lognormal Distributions for Water Intake by Children and Adults

Ann M. Roseberry[1] and David E. Burmaster[1]

Received December 31, 1990; revised July 15, 1991

We fit lognormal distributions to data collected in a national survey for both total water intake and tap water intake by children and adults for these age groups in years: 0 < age < 1; 1 ≤ age < 11; 11 ≤ age < 20; 20 ≤ age < 65; 65 ≤ age; and all people in the survey taken as a single group. These distributions are suitable for use in public health risk assessments.

KEY WORDS: Drinking water ingestion; uncertainty analysis; Monte Carlo simulation; population distributions.

1. INTRODUCTION

When estimating potential public health exposures and risks at or near hazardous waste sites, analysts often assume that children and adults, respectively, ingest 1 and 2 liters of water per day (L/d). These standard assumptions, first published by the Safe Drinking Water Committee of the National Academy of Sciences,[1] now appear routinely in guidance manuals published by the U.S. Environmental Protection Agency (EPA) for the "Superfund" and related programs.[2,3] The EPA also publishes a simple cumulative distribution for drinking water ingestion by adults in the form of a "staircase" function.[3]

Recently, Ershow and Cantor[4] have published a statistical analysis of water intake rates for children and adults in different age groups as measured during and reported by the 1977–1978 Nationwide Food Consumption Survey (NFCS) of the U.S. Department of Agriculture on all foods and beverages consumed during a 3-day period. In their report, Ershow and Cantor divide total water intake into: (i) tap water intake, the sum of water drunk directly as a beverage, and water added to foods and beverages during preparation; and (ii) intrinsic water intake (i.e., the water intrinsic in foods as purchased). Tabulating their results in grams of water per day (g/d, equivalent to the more commonly used volumetric units of milliliters of water per day, ml/d), Ershow and Cantor report binned data and corresponding histograms for total water and tap water intake for these numbers of children and adults in these broad age groups (in years): $n = 403$ for $0 < \text{age} < 1$; $n = 5605$ for $1 \le \text{age} < 11$; $n = 5801$ for $11 \le \text{age} < 20$; $n = 11{,}731$ for $20 \le \text{age} < 65$; $n = 2541$ for $65 \le \text{age}$; and $n = 26{,}081$ for all people in the survey taken as a single group. For each age group, Ershow and Cantor report the data in approximately 12 bins, typically 250 g/d wide for low intake values and 500 g/d wide for higher intake values. Ershow and Cantor also report other aggregations of the data, such as regional differences, not investigated here.

To facilitate quantitative uncertainty analyses in public health risk assessments, we fit lognormal distributions to the binned data for water intake as reported by Ershow and Cantor. We also estimate population-wide distributions for total water and tap water intake based on proportions of people in each age group.

2. METHODS

Defining the variables IR_{tot} and IR_{tap} as the intake rates (in ml/d) of total water and tap water, respectively,

[1] Alceon Corporation, P. O. Box 2669, Harvard Square Station, Cambridge, Massachusetts 02238.

0272-4332/92/0300-0099$06.50/1

we used visualization and exploratory data analysis techniques to view and analyze the binned data.[5–7] Because the histograms and cumulative histograms of the data for each group show long tails to the right, we further investigated histograms and cumulative histograms for the natural logarithms of the binned data. To improve the resolution of our analyses near the origins of several plots, we contacted Abbey G. Ershow (National Heart, Lung, and Blood Institute in Bethesda, Maryland), who graciously supplied additional information in the form of finer binnings of the data for some of the age groups.[8]

Persuaded that Ershow and Cantor's binned data for each age group show the general shape of a lognormal distribution, we used formal estimation techniques to fit lognormal distributions to the cumulative data. More specifically, we fit lognormal distributions of the form:

$$\ln IR \sim N(\mu, \sigma) \quad <=> \quad IR \sim \exp[N(\mu, \sigma)]$$

where IR denotes the intake rate (in ml/d), ln denotes the natural logarithm, exp denotes the exponential function, and $N(\mu, \sigma)$ denotes a normal or Gaussian distribution with mean μ and standard deviation σ.

Following standard statistical techniques for grouped and censored data,[9,10] we estimated μ as the zero-intercept and σ as the slope of the best-fit straight line for ln IR plotted against its z-score. We used ordinary least squares to fit the best straight lines through the transformed values.

2.1. Lognormal Distributions Fitted to the Ershow and Cantor Data

Figures 1–6 show plots of the transformed binned cumulative data and the best fit straight lines for total water (open squares) and tap water (open circles) intake rates for the age groups discussed above: $0 <$ age < 1; $1 \leq$ age < 11; $11 \leq$ age < 20; $20 \leq$ age < 65; $65 \leq$ age; and all people in the survey sample. As seen in these figures, the natural logarithms of the intake rate vary in an approximately linear fashion with respect to the z-scores. In most cases, we note slight deviations from linearity at either extreme. The plots show that the transformed data are generally well fit by straight lines over the range $-3 \leq z \leq 3$ [i.e., $\pm$ 3 standard deviations from the mean (median) of the transformed data, with larger deviations from linearity outside this range].

Table I shows the corresponding estimates for μ and σ, the parameters of the best-fit lognormal distributions, for the total water and tap water intake rates for each group. Table I also shows the R^2 values for the regressions, sometimes called the coefficient of deter-

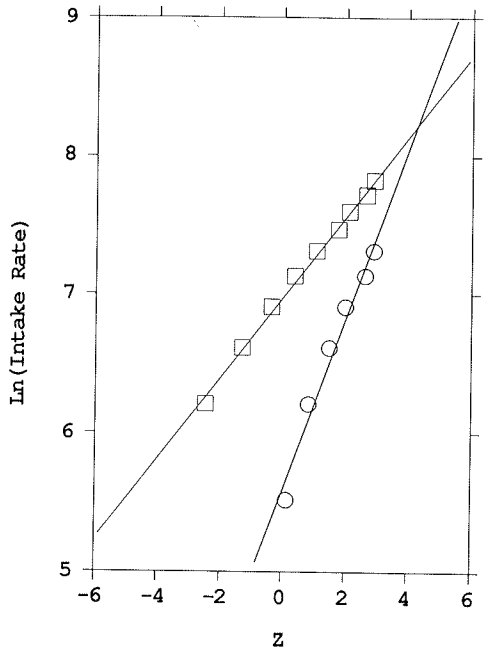

Fig. 1. Distribution of drinking water intake for the age group $0 <$ age < 1: ln(intake rate) vs. z (□ = total water; ○ = tap water).

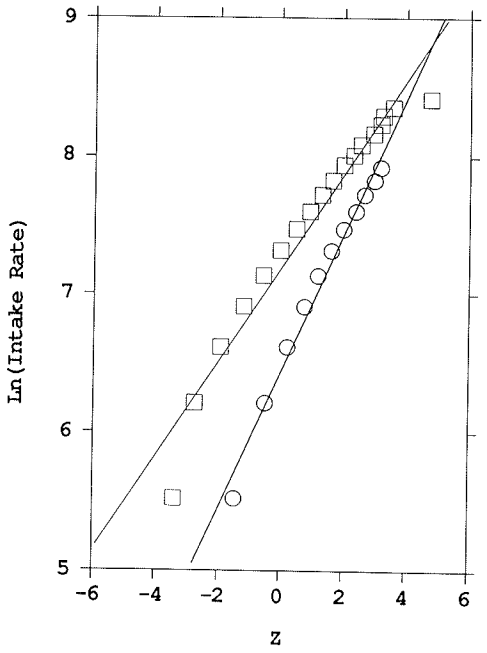

Fig. 2. Distribution of drinking water intake for the age group $1 \leq$ age < 11: ln(intake rate) vs. z (□ = total water; ○ = tap water).

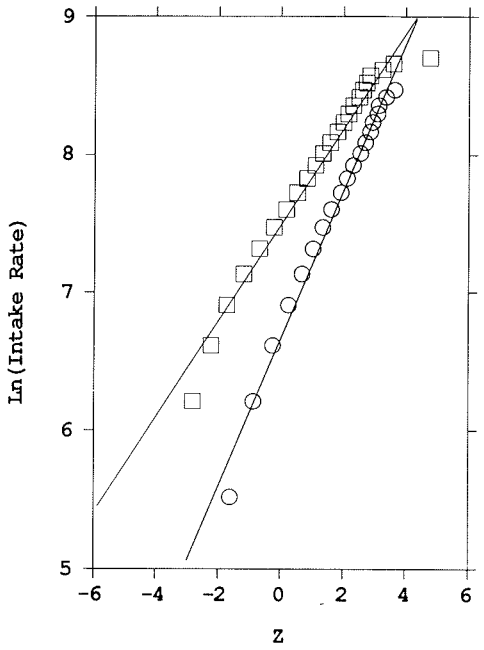

Fig. 3. Distribution of drinking water intake for the age group 11≤age<20: ln(intake rate) vs. z (□ = total water; ○ = tap water).

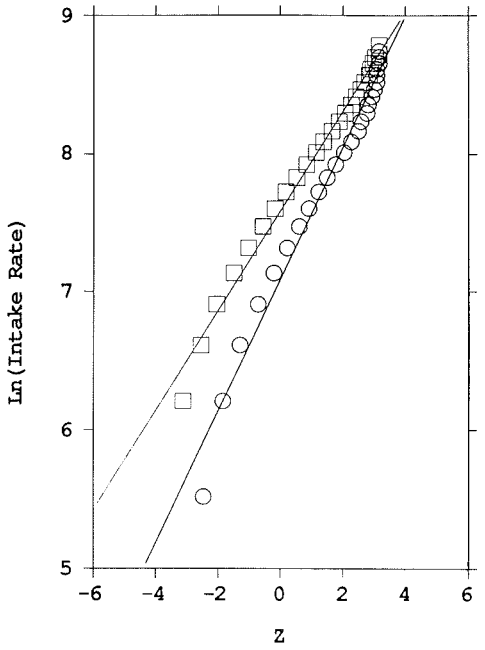

Fig. 5. Distribution of drinking water intake for the age group 65≤age: ln(intake rate) vs. z (□ = total water; ○ = tap water).

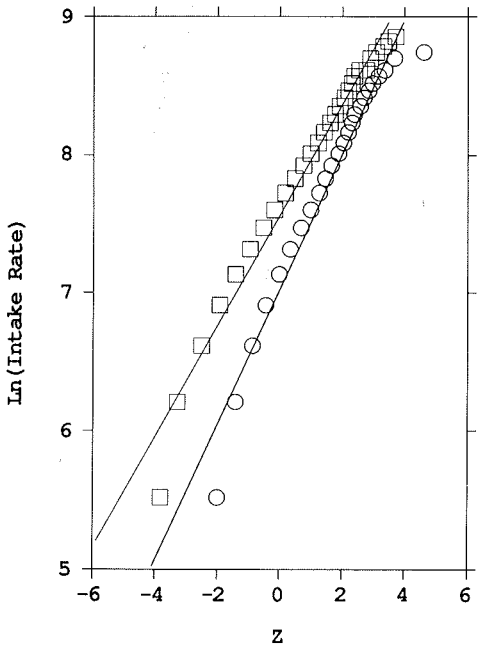

Fig. 4. Distribution of drinking water intake for the age group 20≤age<65: ln(intake rate) vs. z (□ = total water; ○ = tap water).

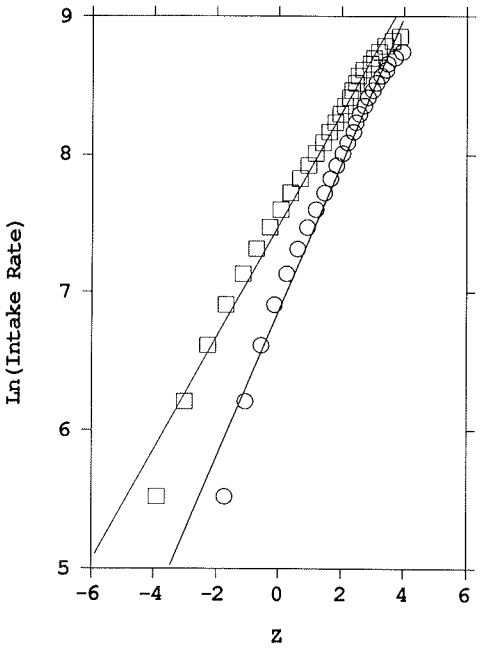

Fig. 6. Distribution of drinking water intake for all people: ln(intake rate) vs. z (□ = total water; ○ = tap water).

Table I. Summary Statistics for Best-Fit Lognormal Distributions for Water Intake Rates

Group	Total water intake		
	μ	σ	R^2
0 < age < 1	6.979	0.291	0.996
1 ≤ age < 11	7.182	0.340	0.953
11 ≤ age < 20	7.490	0.347	0.966
20 ≤ age < 65	7.563	0.400	0.977
65 ≤ age	7.583	0.360	0.988
All NFC survey	7.487	0.405	0.984
Simulated balanced population	7.492	0.407	1.000
	Tap water intake		
0 < age < 1	5.587	0.615	0.970
1 ≤ age < 11	6.429	0.498	0.984
11 ≤ age < 20	6.667	0.535	0.986
20 ≤ age < 65	7.023	0.489	0.956
65 ≤ age	7.088	0.476	0.978
All NFC survey	6.870	0.530	0.978
Simulated balanced population	6.864	0.575	0.995

mination for the lines. All the R^2 values exceed 0.95, implying excellent fits.

Table II shows different quantiles and summary statistics for total water and tap water intake based on these formulae[11]:

$$97.5 \text{ percentile } IR = \exp[\mu + (1.96 \cdot \sigma)]$$
$$75 \text{ percentile } IR = \exp[\mu + (0.6745 \cdot \sigma)]$$
$$50 \text{ percentile } IR = \exp[\mu]$$
$$25 \text{ percentile } IR = \exp[\mu - (0.6745 \cdot \sigma)]$$
$$2.5 \text{ percentile } IR = \exp[\mu - (1.96 \cdot \sigma)]$$
$$\text{arithmetic average } IR = \exp[\mu + 0.5 \cdot \sigma^2]$$

Thus, Table II shows the the $q_{0.25}$ to $q_{0.75}$ range (the interquartile range) and the $q_{0.025}$ to $q_{0.975}$ (the central 95% of the distribution), two ranges of strong regulatory interest.

2.2. Lognormal Distributions for the General Population

The age groups in the NFCS population do not represent the same fractions of the total as the corresponding age groups in the overall national population. More specifically, the NFCS had these approximate fractions of people in the different age groups—0.015, 0.215, 0.222, 0.450, and 0.097, in ascending order—but the overall population has these fractions in the same age groups—0.016, 0.146, 0.128, 0.587, and 0.123, also in ascending order—as estimated by the 1988 Census.[12] Thus, the lognormal distribution fit to the composite NFCS data analyzed by Ershow and Cantor does not have the statistical weights for each age group corresponding to 1988 Census data.

To utilize the modified data, we also estimate the distribution parameters by simulating the distributions for each age group and for the total population using Monte Carlo simulation. We simulate distributions for

Table II. Estimated Quantiles and Arithmetic Averages for Water Intake Rates

Group	Total water intake (ml/d)					
	Percentile					
	2.5	25	50	75	97.5	Arithmetic average
0 < age < 1	607	882	1074	1307	1900	1120
1 ≤ age < 11	676	1046	1316	1655	2562	1394
11 ≤ age < 20	907	1417	1790	2262	3534	1901
20 ≤ age < 65	879	1470	1926	2522	4218	2086
65 ≤ age	970	1541	1965	2504	3978	2096
All NFC survey	807	1358	1785	2345	3947	1937
Simulated balanced population	808	1363	1794	2360	3983	1949
	Tap water intake (ml/d)					
0 < age < 1	80	176	267	404	891	323
1 ≤ age < 11	233	443	620	867	1644	701
11 ≤ age < 20	275	548	786	1128	2243	907
20 ≤ age < 65	430	807	1122	1561	2926	1265
65 ≤ age	471	869	1198	1651	3044	1341
All NFC survey	341	674	963	1377	2721	1108
Simulated balanced population	310	649	957	1411	2954	1129

the total population by summing contributions from the individual age groups using a random sample with appropriate contributions for the age groups as estimated for the United States in the 1988 Census.[12] The older age groups in the NFCS population were not balanced by gender, a difficulty we could not correct in the present analysis.

Using Crystal Ball,[13] a Monte Carlo simulation program with latin hypercube sampling, we simulated a total of 6000 draws from the distributions for the separate age groups to estimate distributions for total water and tap water intake for the (properly age-balanced) U.S. population. That is, in 6000 draws, Crystal Ball drew from the distributions for the different age groups in proportion to the age distribution of the national population in 1988.

Figure 7 shows the results of the 1988-population-balanced simulation—with the values transformed, plotted, and regressed as were survey values in the first six figures (with filled symbols). Even though this simulation produces a mixed statistical population by definition, the results appear lognormally distributed. From the regressions, we estimate μ as 7.492 and 6.864 and σ as 0.407 and 0.575 for total water and tap water intake, respectively. The regressions have $R^2 = 1.000$ and 0.995, respectively. Table II shows the corresponding quantile values and summary statistics estimated from the simulations.

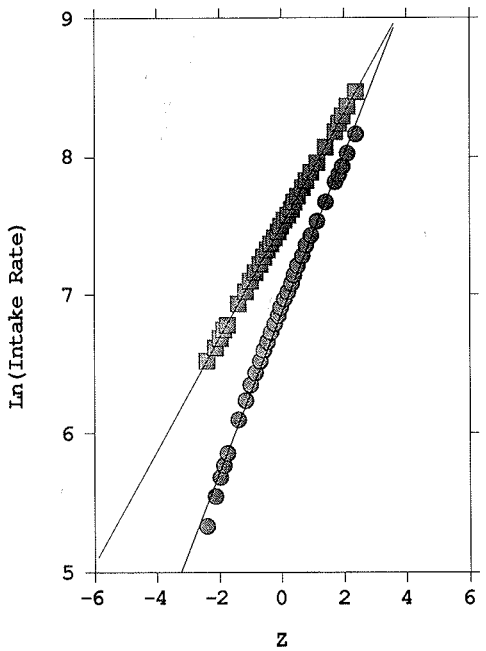

Fig. 7. Distribution of drinking water intake for the simulated balanced population: ln(intake rate) vs. z (■ = total water; ● = tap water).

3. DISCUSSION

As shown in Figs. 1–6, plotting the natural logarithm of the intake rates vs. z produces linear results for all data sets. Linear regression gives R^2 values ranging from 0.953–0.996. We attribute lower R^2 values for some data sets to deviations from linearity at the extremes. These deviations from linearity are most apparent in the 11–20 age group for total water ingestion. Limiting the data to within $\pm$ 2 standard deviations improves R^2 values. For the 11–20 age group for total water ingestion, such limitation results in an R^2 value of 0.999, compared to 0.966 for the total data set. This linearity of the natural logarithm of the intake rates vs. z-scores provides confidence that the data are lognormally distributed.

Tap water intake is a subset of total water intake; thus, total water intake must always be greater than tap water intake. However, people who drink greater quantities of tap water do not necessarily ingest greater quantities of food. That is, the ratio of total water to tap water is not a constant value. Because the tap water intake rates have lower medians but higher standard deviations than do the total water intake rates, the regression lines in Figs. 1–6 intersect at high values outside the range of real intake values.

Although restriction of the data to $\pm$ 2 standard deviations generally improves the fit between the data and the lognormal distribution, we estimate distribution parameters using all the data. While it is important to recognize deviations from the ideal at higher volumes, we cannot justify exclusion of any of the data.

Finally, simulations with our estimated parameters match the original data with good agreement. That is, constructing ideal lognormal distributions from the estimated means and standard deviations presented in Table I gives percentile values very similar to the observed percentile values presented in Table II.

4. CONCLUSIONS

We fit lognormal distributions to data for total water and tap water intake as reported by the Nationwide Food Consumption Survey for several age groups. To estimate lognormal distributions for the overall population, we used Monte Carlo simulations to draw from the appropriate lognormal distributions for the different age groups.

Overall, the robust results for the different age groups

and for the simulated total population provide lognormal distributions for tap water intake suitable for use in public health risk assessments such as those performed and required by the EPA.

ACKNOWLEDGMENTS

We thank Abbey G. Ershow of the National Heart, Lung, and Blood Institute, Bethesda, Maryland, and Kenneth P. Cantor of the National Cancer Institute, Bethesda, Maryland, for providing us with additional data from their study. We also thank William M. Rand of the Tufts University Department of Community Health in Boston, Massachusetts, for his help with the statistical analyses of the grouped data. Alceon Corporation funded this research.

REFERENCES

1. National Academy of Sciences, *Drinking Water and Health* (Report by the Safe Drinking Water Committee, National Academy of Sciences, Washington, D.C., 1977).
2. U.S. Environmental Protection Agency, *Risk Assessment Guidance for Superfund, Vol. 1, Human Health Evaluation Manual, Part A* (Interim final report, Office of Emergency And Remedial Response, EPA/540/1-89/002, Washington, D.C., December 1989).
3. U.S. Environmental Protection Agency, *Exposure Factors Handbook* (Office of Health and Environmental Assessment, EPA/600/8-89/043, Washington, D.C., July 1989).
4. A. G. Ershow and K. P. Cantor, *Total Water and Tapwater Intake in the United States: Population-Based Estimates of Quantities and Sources* (Life Sciences Research Office, Federation of American Societies for Experimental Biology, Bethesda, May 1989).
5. J. W. Tukey, *Exploratory Data Analysis* (Addison-Wesley, Reading, Massachusetts, 1977).
6. J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey, *Graphical Methods for Data Analysis* (Wadsworth International Group, Belmont, California, and Duxbury Press, Boston, Massachusetts, 1983).
7. SYSTAT, Inc., *Users Manuals for SYGRAPH Graphics and SYSTAT Statistics* (Evanston, Illinois, 1990).
8. A. G. Ershow, (personal communication from researcher at the National Heart, Lung, and Blood Institute in Bethesda, Maryland, 1990).
9. C. C. Travis and M. Land, "Estimating the Mean of Data Sets with Nondetectable Values," *Environmental Science and Technology* **24,** 961–962 (1990).
10. W. M. Rand, (personal communication from professor of statistics at Tufts University's Department of Community Health, Boston, Massachusetts, 1990).
11. N. A. J. Hastings and J. B. Peacock, *Statistical Distributions: A Handbook for Students and Practitioners* (Butterworth and Company, London, England, 1974).
12. U.S. Department of Commerce (Bureau of the Census, Statistical Abstract of the United States for 1990, 110th ed., 1990) p. 12.
13. Market Engineering Corporation, *Crystal Ball User's Guide for Release 2* (Denver, Colorado, 1990).